

DRAFT General statement of work for procurement of software and consulting services in development of several transactional databases, a data warehouse, and supporting web-enabled online analytical processing.

This solicitation notice is for comments and information related to draft requirements for development of a pilot data management project associated with the Ecological Synthesis Team of the National Water Quality Assessment Program (NAWQA) of the U.S. Geological Survey (USGS). Desired products of the pilot effort include a transactional database for NAWQA ecological data, a data warehouse for this ecological data, and web-enabled analytical capabilities for using the data warehouse.

The overall objectives include: (1) development of a pilot transactional database implemented in three NAWQA study units which is populated by the study units using computer entry forms, data from existing databases, and other data sources; (2) development of a pilot data warehouse that draws on data from the three transactional databases; (3) development of web-enabled online analytical processing tools; (4) demonstration of the functionality of the input, storage, and output systems and the integrity of the data throughout the systems. A schematic of the Ecological Synthesis data management model is shown in figure 1. This is a pilot effort; if successful, a decision could be made to do a competitive procurement for similar services in a larger USGS application. It is desirable that the pilot be instituted through a rapid, incremental development approach and that all databases be implemented within an off-the-shelf- relational database management software systems such as Access, Oracle, or INGRES.

A. Background

The US Geological Survey's National Water-Quality Assessment (NAWQA) Program collects physical, chemical, and biological data to assess the condition of, factors controlling, and trends in the quality of the Nation's surface- and groundwater resources. Data are collected by 59 Study-Unit teams, which are located in District Offices distributed across the contiguous United States, Alaska, and Hawaii (fig 2). Study-Unit teams rotate between periods of high intensity (3 years) and low intensity (6 years) data collection activities with approximately one-third of the teams involved in high intensity data collection during any one year. Study-Unit Team staffing levels are adjusted to reflect the intensity of sample collection activities. Data collected by Study-Unit teams are either processed locally and entered into local (District) databases or data are generated at the U.S. Geological Survey's National Water Quality Laboratory (NWQL, Arvada, CO) and electronically transferred to District databases. Each District database represents part of the U.S. Geological Survey's National Water Information System (NWIS). Certain data (fish, algae, and invertebrate community data) are not currently incorporated in NWIS and are stored by Districts in a variety of electronic forms. These data are generated both by the Study-Unit Team and by the NWQL.

In addition to the Study-Unit Teams associated with USGS Districts, the NAWQA Program has five National Synthesis groups with members located in Sacramento, CA; Raleigh, NC; Atlanta, GA; Reston, VA; and Rapid City, SD). These National Synthesis groups both supply supporting

information to the Study-Unit Teams (e.g., national land-use information) and aggregate data from Study-Unit databases. Data aggregated by National Synthesis groups are made available to Study-Unit Teams, the USGS, other Federal and State agencies, and the general public. It is desired that these data be distributed using queriable databases accessible through the World Wide Web.

B. NAWQA Ecological Synthesis Transactional Database

A pilot NAWQA Ecological Synthesis transactional database will be developed that contains site information and data related to chemical and biological samples, sample results, and taxonomic information. The database system to be developed will also include electronic field forms that will enable data to be entered on a laptop computer and transferred into the relational database. This transactional database will be developed and implemented in three NAWQA study units, located in Raleigh, NC (Albemarle-Pamlico), Denver, CO (Upper Colorado River), and Baton Rouge, LA (Acadia). The relational database will be implemented using ACCESS as the database engine. These three pilot databases will be utilized in a data warehouse system, described in Section C. A general schematic of the transactional database is contained in figure 3.

1. Site information

There are a number of sources of information that describe the locations, or sites, where surface water quality data are collected for the NAWQA program, and the drainage basins associated with the sampling sites. For the purpose of this document, the phrase site information will include information about both the actual sampling location and the upstream drainage basin, unless otherwise noted.

This site information is maintained within the National Water Information System database (NWIS) system in USGS District offices, within local project databases in District offices on databases other than NWIS, and in databases located at NAWQA National Synthesis offices in Oklahoma City, OK and Sacramento, CA. District-level NWIS data are maintained in a non-relational database residing within an INGRES software system on a Data General UNIX platform. Other site information is maintained in data formats including Excel, Access, SAS, and RDB.

These databases are not currently synchronized and may contain site information that is inconsistent. Because many data queries will depend upon information stored in this site characteristic file, there is a need to develop a single NAWQA site characteristics data system that makes use of these data, identifies inconsistencies, allows users to understand which data sources are inconsistent, and enables users to reconcile these differences and store site data in “master” site table(s) within the transactional database.

The site tables portion of the ecological synthesis transactional database will contain site information for three NAWQA study units, located in Raleigh, NC (ALBE, a 1991 study unit), Denver, CO (UCOL, a 1994 study unit), and Baton Rouge, LA (ACAD, a 1997 study unit). A transactional database will reside in each location. These transactional databases shall be accessed electronically by a central data warehouse, located in Raleigh, NC or Denver, CO.

- 1) The site table in the transactional database shall be populated initially from site data that has been aggregated from NAWQA 1991 study units into databases located Sacramento, CA.
- 2) The local transactional database will have the ability to retrieve site characteristic data from NWIS and from NAWQA project level databases that are maintained by individual project personnel at each of the three study unit locations described above. Data are created at a variety of spatial scales, including the study unit, a basin or watershed, a segment, a reach, and a transect. Non-NWIS databases are usually maintained in MS-Access or MS-Excel on a PC platform, or in ARC/INFO datafiles on a UNIX or Windows NT platform
 - a. The system shall have the ability to load data from these existing local site and basin databases, including habitat data and other site data from electronic field forms.
 - b. The database shall have the capability to load both column-formatted and delimited text files, including files that have a site identifier in the first column and basin or site data in the other columns. Users shall be able to define the names and data structure of these text files in a way that the information can be automatically read into the existing site tables.
 - c. The system shall have the capability to retrieve site-related data entered using electronic field forms, either directly from a laptop computer or from an electronic file. Field form variables are described in a spreadsheet located at <http://sg1dncrlg.er.usgs.gov/nceg/ESPT/params.xls>.
 - d. The system shall provide the capability to identify and list inconsistencies in station identifiers, station names and all other site characteristic variables in all databases from which site information is drawn, and notify administrators of the local databases. The system shall make use of fuzzy logic or other information matching capabilities to group and identify sites that are closely related in terms of station identification and station name.
 - e. The systems shall enable the user to use query forms to help populate the database, with the user specifying a station identifier and the system identifying related variables and their values in the NWIS database, and other local databases for potential inclusion in the database.
- 3) The system shall notify the data base user and administrator of inconsistencies that arise during initial data loading, data refreshing, and data retrievals associated with the site table(s).
 - a. The data base administrator will have the ability to change the database to reflect updated information related to these inconsistencies.
- 4) The system shall have the ability, for user-designated sites, to enable users to select site variables from a pick list and create tabular and graphical summaries of the site data.
 - a. The database system shall have the capability to produce reports of site information to the computer screen or to electronic files, with report criteria being set by users making use of SQL queries on any variables in the site tables and conventional Boolean logic.
 - b. The system shall have all capabilities described in section C1b, below, for selecting sites.
 - c. Column and row identifiers shall remain visible when the user scrolls down a data table on the computer screen.
 - d. Each data table shall begin with a column listing station identifiers that can be linked to more detailed station information.

- e. The system shall identify one variable for each column in a data array (i.e., each column shall list a parameter name or a modifier of the parameter value).

2. Electronic field forms

There is a general need, on the part of biologists working with individual NAWQA studies, to enter data collected during sampling trips. Data capture requirements are specified in protocols that have been developed for habitat characterization, and for fish, algae, and invertebrate sampling.

- 1) Electronic field forms (EFF) shall be developed that allow entry of data required in biological protocol documents (habitat, fish, algae, and invertebrates) on a laptop computer into an Access database.
 - a. Data entry tools for habitat shall reflect evaluation of Ecotools, an electronic field form for habitat data, described at <http://sgl1dnclrg.er.usgs.gov/nceg/ecotools/index.html>.
 - b. Data entry tools shall reflect evaluation of electronic field forms under development by Frank Crenshaw and Bob Boulger of the Colorado District.
 - c. The EFF shall include a remark field for the overall quality of the sample with a domain including the values “poor” “fair” “good” and excellent”. There will also be a related comment field to explain these remark codes.
 1. This remark field will be used, for example, if data appear to be erroneous, or the day the sample was taken was unusual circumstances (e.g., higher than usual flow, water was turbid and they couldn’t see well that day), or the study unit biologist knows that sample collection procedures deviated substantially from protocols. A remark will alert the data analyst that a decision is needed as to whether to include this data point(s) in analyses.
 2. This field could apply for a single sample (a result) or a group of results (e.g., DTH, RTH, QMH algal sample or RTH, QMH invertebrate sample)
- 2) Electronic field forms shall allow entry of all data needed to generate a laboratory Analytical Services Request (ASR) form.
- 3) The EFF shall have the capability to link to the site table(s) in the transactional database, and import all site-related information required by the protocol and already existing in the site table(s).
- 4) The EFF shall have the capability to automatically check the domain of variables while data are being entered, to allow biologists to indicate the units of data being entered, and automatically convert these data into units required by the protocol.
- 5) The EFF shall have the capability to link into related taxonomic tables needed to fill out the field forms, including AFS tables for fish.
 - a. The EFF user shall be able to begin typing a fish species name and automatically see names with that spelling from the AFS fish species list, and choose a name from this list to populate the field form.

- 6) Electronic field forms shall have the capability of generating hard-copy reports and ASRs, including reports of the value of all variables entered in the EFF, in the units that values were entered.
- 7) The EFF shall have the capability to transmit analytical service request forms (ASRs) for both chemical and biological samples via the World Wide Web (WWW) to the National Water Quality Laboratory in machine-readable format.
- 8) The design and structure of the Access database associated with the EFF shall be consistent with the design and structure of the transactional database described in subsection B3 below.
- 9) The transactional database described in B3 shall be able to automatically upload data from the EFF Access database.

3. Biological and chemical samples results and taxonomic tables

The Ecological Synthesis transactional database will also store data tables related to biological and chemical samples, results, and taxonomic tables. Some of this data originates from NWIS databases located in each of the three Districts that are participating in the pilot project. Other data are derived from project databases that are not part of NWIS. Additional data are created by study unit scientists either using electronic field forms or hand input into spreadsheets. Finally, the Biological Unit of the NWQL in Arvada, CO creates data. In general, the transactional database shall have the capability to electronically load data from these sources and have the capability to scan these databases for updated or revised data and reload these new data.

- 1) A database structure will be created that includes variables described in a spreadsheet located at <http://sgl1dnclrg.er.usgs.gov/nceg/ESPT/params.xls>.
- 2) The database structure will include all information contained in the BDAS data tables, described in Appendix C of the BDAS documentation available <http://sgl1dnclrg.er.usgs.gov/nceg/bdas/index.html>
- 3) The data structure shall be consistent with the general data model shown in figure 3.
- 4) The transactional database for the Raleigh, NC site will be populated initially using chemical, streamflow, habitat, fish, algae, and invertebrate data for “basic fixed sites” of the ALBE study residing in a pilot data warehouse currently located in the Nutrient National Synthesis project office in Denver, CO. The structure of this data warehouse is shown in Figure 5.
- 5) The system shall include a plan for populating the Denver (UCOL study) and Baton Rouge (ACAD study) transactional databases with data from all surface water chemical and biological sampling sites, including synoptic sampling sites.
- 6) The database shall enable the user to store data collected by continuous temperature monitors such as the HOBO temperature monitors, including daily mean, min, max, annual degree days.

- 7) The system shall enable the user to electronically retrieve data from NAWQA project level databases that are maintained by individual project personnel at the study unit locations described above. These databases are usually maintained in MS-Access, MS-Excel, SAS, or RDB on a PC or UNIX platform.
- 8) The local transactional database shall enable to local user to electronically retrieve data from the local District NWIS database, including ADAPs, GWSI, and QWDATA.
 - a. The user shall be able to use menus and pick lists to indicate the criteria for data retrieval from NWIS into the transactional database, including parameter names and time periods.
- 9) All retrievals, queries, downloads, and uploads that require site information shall make use of the site tables in the transactional database.
- 10) The database shall be able to electronically download and upload chemical and biological data into the National Water Quality Laboratory Laboratory Information Management System (LIMS) and related lab databases, using an electronic ASR and electronic queries.
 - a. The system shall have the ability to retrieve tab-delimited data related to invertebrates (http://wwwnwql.cr.usgs.gov/USGS/fmt_invert.html), and algae, and all related taxonomic tables.
 - b. The taxonomic table at the National Water Quality Lab shall be the guiding table, and taxonomic tables in local transactional databases shall be automatically synchronized with the master tables at the lab.
- 11) The database system shall enable users to have several methods of getting raw data into the results database:
 - a. The system shall include electronic field forms that allow a user to enter data using electronic data entry tables on a laptop computer.
 - 1) These forms should QA the data, as it is being input, and give the user helpful error messages to resolve “disallowed” values.
 - b. The system shall allow data to be uploaded from databases populated by use of electronic forms, including electronic field forms.
 - c. The system shall have the ability to import whole QA’d field form or habitat datasets (such as the SU91’s) in a batch mode.
 - d. The system shall have the ability to interactively import unQA’d datasets (such as the datasets that SU94’s might have, but have not yet gone through a formal process of national QA) in a batch mode.

4. Report creation capabilities

The transactional database shall allow the user at the District level to query the database and produce data reports.

- 1) The system shall enable sites used in data queries to be identified using Boolean queries of variables contained in the site table(s) and selection of sites using an image map, as described in Section C1b, below.

- 2) The system shall allow users to access both “raw” data and various data summaries, as described in Section C2 below.

C. Data Warehousing capabilities desired by users

Data warehousing requirements are based on the needs of USGS database users (internally and externally) to access, query, and analyze data collected by the USGS as part of the National Water Quality Assessment program (NAWQA). In this pilot project, a data warehouse shall be located in the USGS District office in either Raleigh, NC or Denver, CO. This data warehouse shall be able to compile data from the three pilot transactional databases, allow the users to produce reports, and allow users to download data files. The data warehouse development tasks will include reviewing and evaluating existing data warehousing capabilities that exist within the NAWQA program, and making recommendations about incorporating this design into the pilot effort. The data warehouse system shall be implemented in a commercial, off-the-shelf relational database system such as Oracle or Ingres.

- 1) The data warehouse system shall have the ability to make repeated selections of data from the three pilot transactional databases based on user specified criteria for sites, attributes, and analytic data characteristics.
 - a. The system must be able to obtain data from nationally distributed transactional databases.
 1. The system shall have the ability to selectively retrieve and update data from the three pilot transactional databases, located in Raleigh, NC, Denver, CO, and Baton Rouge, LA
 - a) The system shall have the capacity to retrieve environmental and quality control sample data associated with basic fixed sites and synoptic sampling sites that are stored in the transactional database.
 - b) The system shall be able to load only changes into the warehouse that have been made since the last time the data warehouse was refreshed.
 - c) Data aggregation from Study Unit to National should be automatic with a minimum of user intervention.
 2. The data warehouse shall be loaded, initially, with water quality and streamflow data from the NAWQA 1991 study units that have been loaded into a data warehouse in Denver (see Fig. 2), with 1991 study unit biological data from the National Water Quality Laboratory (NWQL), and with habitat, algae, fish, and invertebrate data from the Denver National Synthesis database.
 3. The system shall include the capability to automatically run quality assurance algorithms that search and find outlying or missing values and offer users the ability to impute, correct, or leave unchanged the values. The data warehouse should have the ability to automatically aggregate data from Study-Unit databases and other national synthesis databases into the ecological national synthesis data warehouse.

- b. The system shall enable users to select water quality and streamflow data for sampling locations designated by users. The designations shall be based on user-specified criteria about site-attributes, sample-attribute, or streamflow criteria.
 1. The system shall enable users to select one or more sites using the following site attributes: study unit name; state name; county name; hydrologic unit code (HUC); Geologic unit name and code; site name and identification number; time period; and latitude-longitude location. A pick list should be available for all elements other than latitude-longitude, and the users shall be able to specify more than one value from any or all of the pick lists (e.g., sites from counties A and B, for a designated time period, in specified geologic units).
 2. The system shall enable users to select one or more sites using an image map. Examples of such site querying capabilities are contained in: EPA Envirofacts Warehouse National SITEINFO Request Form (<http://www.epa.gov/cgi-bin/enviro/siteinfo.pl>), EPA Region 10 SITEINFO Request Form (<http://www.epa.gov/region10/gisapps/siteplus.html>), and EPA Envirofacts Query Mapper (<http://www.epa.gov/enviro/html/multisystem.html>).
 3. The system will allow users to select one or more sites based on user-specified hydrologic conditions (e.g. all sites with mean daily discharge larger than a specified amount).
 4. The system will enable users to select water-quality samples for selected sites by sample attribute-criteria, including sample date and time, sample medium type, sample constituent group (e.g. all nutrients, all alkalinity), and sample purpose codes.
 5. The system shall enable the user to specify selection criteria for discrete and continuous attributes using the following Boolean qualifiers: and, or, in, not in, between, less than, greater than, equal to, not equal to, less than or equal to, greater than or equal to, and like.
 - a) The system shall enable the user to specify more than one Boolean-qualifier set for each attribute.
 - b) The system shall enable users to select sites by multiple attributes, such as: site numbers, study unit names, county codes, state codes, HUCs, GEOL Unit codes, and latitude-longitude.
 - c) The system shall enable the user to select samples based on a range of sample dates and times.
 - d) The system shall enable the user to select sites based on a range of hydrologic conditions.
 - e) The system shall enable the user to select samples based on Boolean search by presence of one or more parameters picked off lists of parameters.
 6. The systems will enable users to view site selection criteria and/or sample selection criteria, to store these site selection criteria during a session, to edit these criteria, and to download these criteria.

7. The data warehouse should have the ability to automatically aggregate data from Study-Unit databases.
 8. The system shall enable the user to execute a one-step retrieval of sites or samples by multiple combinations of user-specification tools described above.
- 2) The system shall allow users to have access to both “raw” data and various summaries of data from all these databases.
- a. The system shall write data out water quality data in formats compatible with commonly used software packages in the USGS, including ARC/INFO, CANOCO, and SAS, and in tab-delimited format.
 - b. The system shall summarize and write out streamflow data in a format compatible with Indicators of Hydrologic Alteration (IHA) software (Contact Chuck Smythe for information about this format, csmythe@ossinc.net).
 - c. The system shall have the capability to allow users to pick water quality variables and summarization options for one or more stations from a pick list and produce tabular and graphical summaries of data.
 1. Tabular data manipulation options should include parametric and non-parametric measures of central tendency and dispersion, ordination, and cluster analysis.
 2. Tabular and graphical summaries developed in the data warehouse shall be produced using software such as SAS, Excel, CANOCO, and Twinspan.
 3. For fish, invertebrate, and algae data at any site, the system shall be able to compute and report the following measures that may be chosen by users: fish (Taxa richness, Abundance, percent taxonomic groups, Percent tolerant fish, Percent omnivorous fish, Percent non-native fish); invertebrates (Taxa Richness, Taxonomic groupings, Functional groups, tolerance, Abundance, Taxonomic groupings, functional groups, Dominance, Tolerance); and algae (Biomass (biovolume), Taxa richness, Abundance, Autecological guilds {Nitrogen-fixers, Oligotrophic and oligothermal, Eutrophic, Facultative nitrogen-heterotrophic, Halophilic (salt), Siltation tolerant, Cosmopolitan})
 4. The system shall have the ability to choose variables and pick graphical summarization techniques, including box plots, comparison box plots across categories of nominal scaled variables, Lowess and other data smooths, similarity trees, cluster analysis, ordination figures, and three-dimensional data plots.
 5. The system shall have the capability to allow users to choose and graph transformations of quantitative data, including base 10 and natural logarithms.
 6. The system shall allow interactive rotation of data plots.
 - d. The system shall enable users to write out these summaries to a text file and to print these summaries to a local printer.
 - e. The system should have the ability to aggregate site information and physical and chemical data according to user specifications about: sites to be included in the retrieval, environmental and quality control variables to be retrieved, daily value streamflow records, and the time period covered by the retrieval.

1. The system shall enable the user to specify sites using the methods described in section C1b above.
2. The system shall enable the user to specify a retrieval of all water quality or streamflow data for one or more stations, a subset of data based on user-defined constraints, or a subset of data in the form of a summary of user-specified data characteristics.
 - a) User-defined constraints can be given in terms of data for a certain time period, data whose values correspond to user-specified Boolean search constraints.
- f. The system shall identify columns of data in tabular outputs to the screen or files to be downloaded.
 - a) Column identifiers shall remain visible when the user scrolls down a data table on the computer screen.
 - b) Each data table shall begin with a column listing a station identifiers that can be linked to more detailed station information contained in the site characteristics database described in Section B, below.
 - c) The system shall identify one variable for each column in a data array (i.e., each column shall list a parameter name or a modifier of the parameter value).
 - d) The system shall provide output with a measured data value and a remark code.
- g. The access to data shall be read-only.
- h. The systems shall provide a firewall between the data warehouse and WRD databases and applications.
- i. The system shall provide a scaleable hardware architecture that can be expanded to meet future use demands.
- j. The systems shall enable users to execute desired tasks easily and efficiently.
 1. The system shall enable the user to select and retrieve data from one centralized web page, without requiring the user to know the administrative structure of USGS or the locations of specific databases.
 2. The system shall enable users to perform a typical scenario of specification of selection criteria, retrieval, and output using fewer than 5-7 web pages.
 3. The system shall enable users to perform a typical scenario of specification of selection criteria, retrieval, and output in less than five minutes.
 4. The system shall execute retrievals of 10,000 lines of data within a maximum of one minutes during times of typical system load, and shall indicate to the user the time remaining before completion of the retrieval.
 5. The system shall enable users to retrieve data without training or a manual.

D. Development of a Web-enabled online data processing system

1. The system shall produce reports that reflect the functionality of BDAS (Biological Data Analysis System) into the ecological database (e.g., ability to aggregate abundance data at higher taxonomic levels, diversity indices, ability to attach aggregation characteristics to specific taxa and to modify these characteristics on a site by site basis).
 - a. A description of BDAS and its functionality can be obtained at <http://sgl1dnclg.er.usgs.gov/nceg/bdas/index.html>.
2. The web-enabled application should have an ability to link data in the data warehouse with certain application software, including ARC/INFO and SAS.
 - a. The systems shall be able to run ARC/INFO, SAS, CANOCO, and Twinspan programs and produce tabular and graphical content for passive viewing on the client, typically with a Web browser, and for downloading. Common types of content will include development of base maps indicating locations of sampling sites and other base map information, statistical summaries of data, graphical summaries, and a limited set of common analytical procedures for user-designated constituents, including measures of central tendency and dispersion and flow adjusted and non-flow adjusted trend analysis.
3. The web application should have the ability to create and distribute reports created by simple client queries, passed through the web server to a data repository, formatted as a report to be viewed on a web browser or downloaded to the client.
4. The web-application should allow the user to “drill down” in their queries, changing direction in the query as new information is revealed.
 - a. For example, a user should be allowed to pursue a series of queries that might develop as the data are explored. An analysts might look at several temporal scales of chemical water quality data (monthly, quarterly, annual) and compare these exposures to an annual fish response, as measured by several response variables, at all sites in a study unit. This same analysis might be expanded to look at all sampling sites within an ecoregion or set of study units, or for all sites with a characteristics set of average or extreme streamflow conditions. Finally, the analyst might look only at a single response characteristic, for basins that have designated land use conditions. All of these queries can help the analyst discover patterns in the data and focus additional analysis or further sampling efforts.

**NAWQA Transactional Database
(in each Study Unit)**

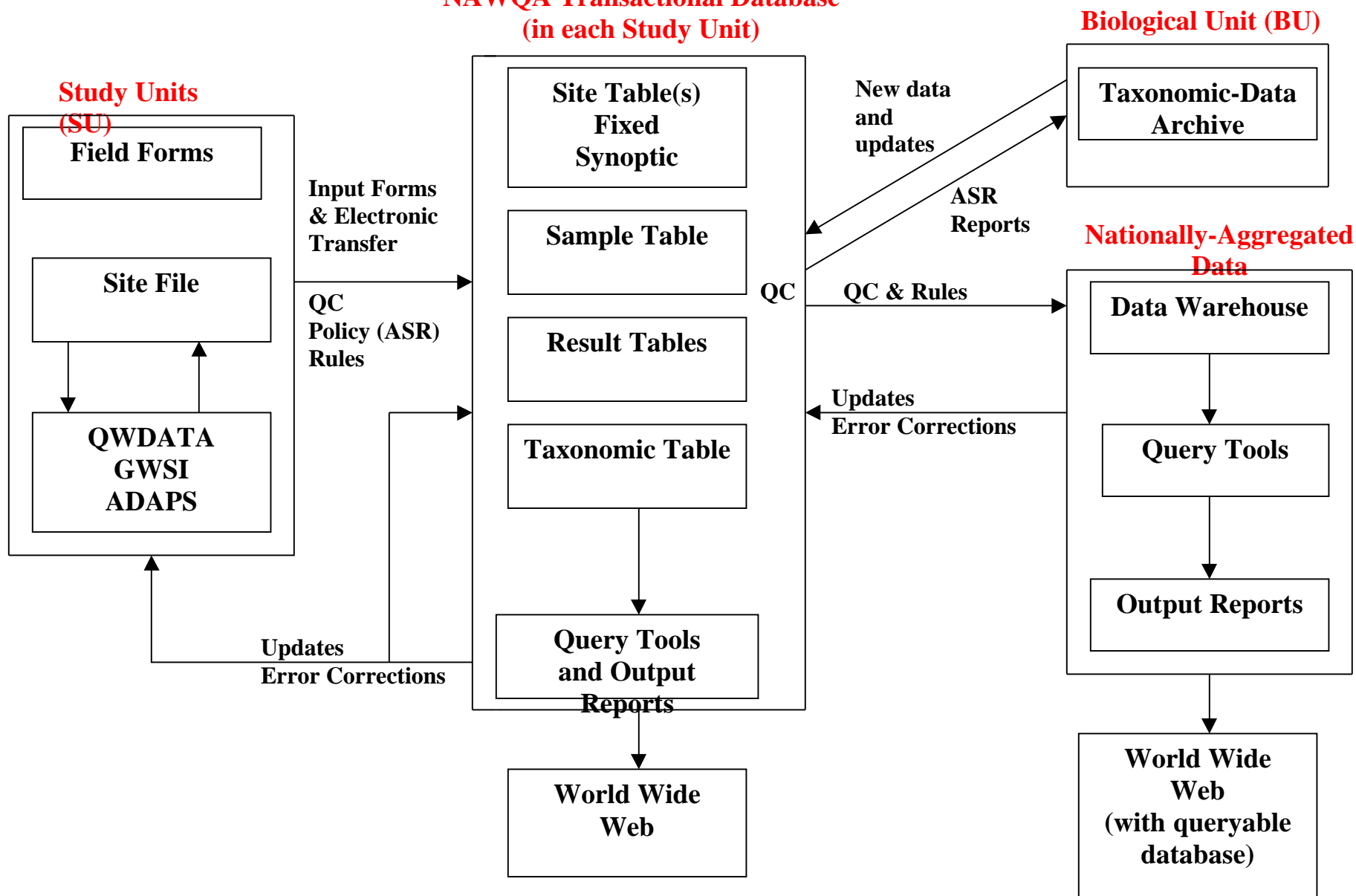
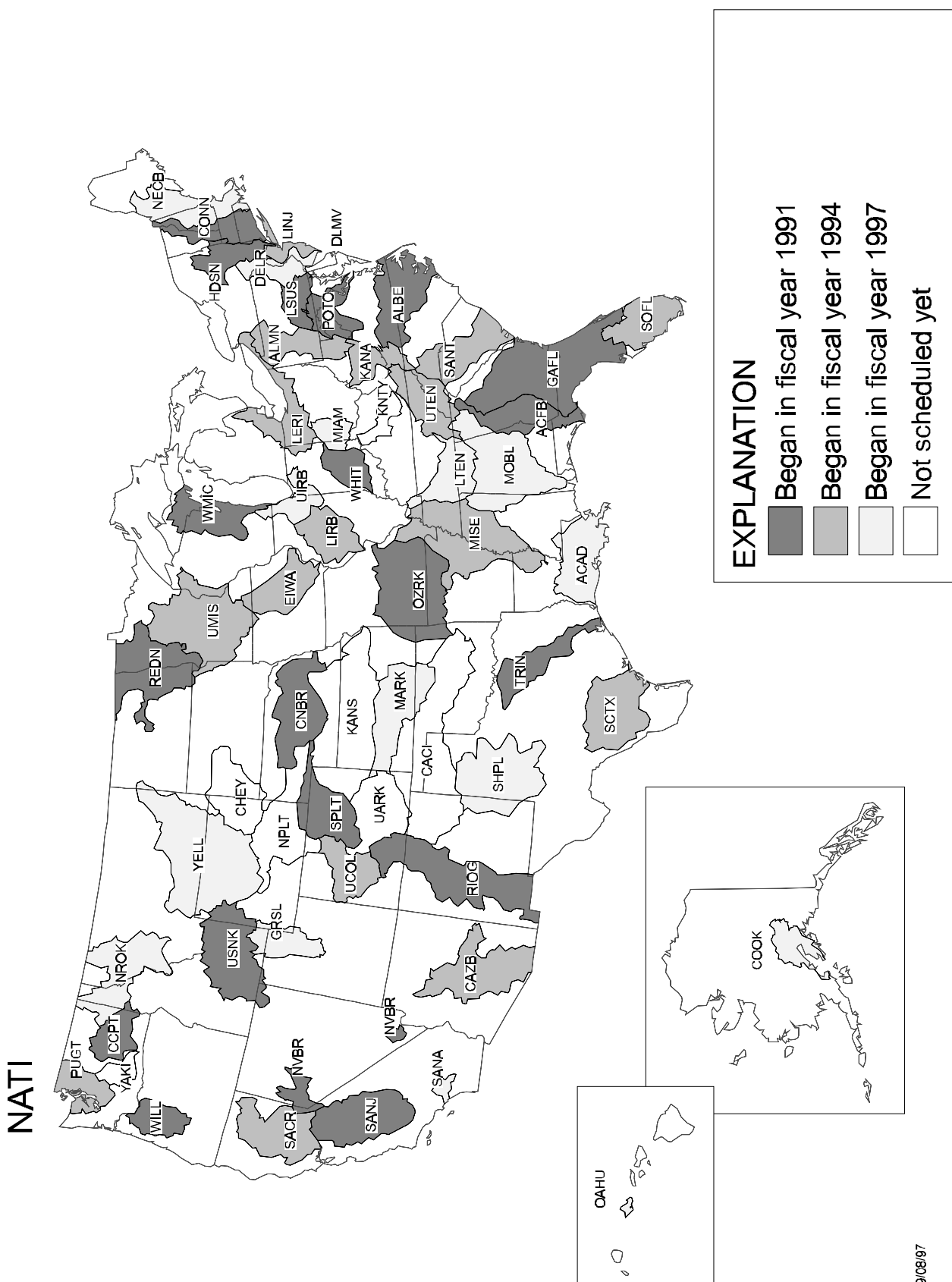


Figure 1 ECOLOGICAL SYNTHESIS DATA-MANAGEMENT MODEL

Fig 2. National Water Quality Assessment Program Study Units



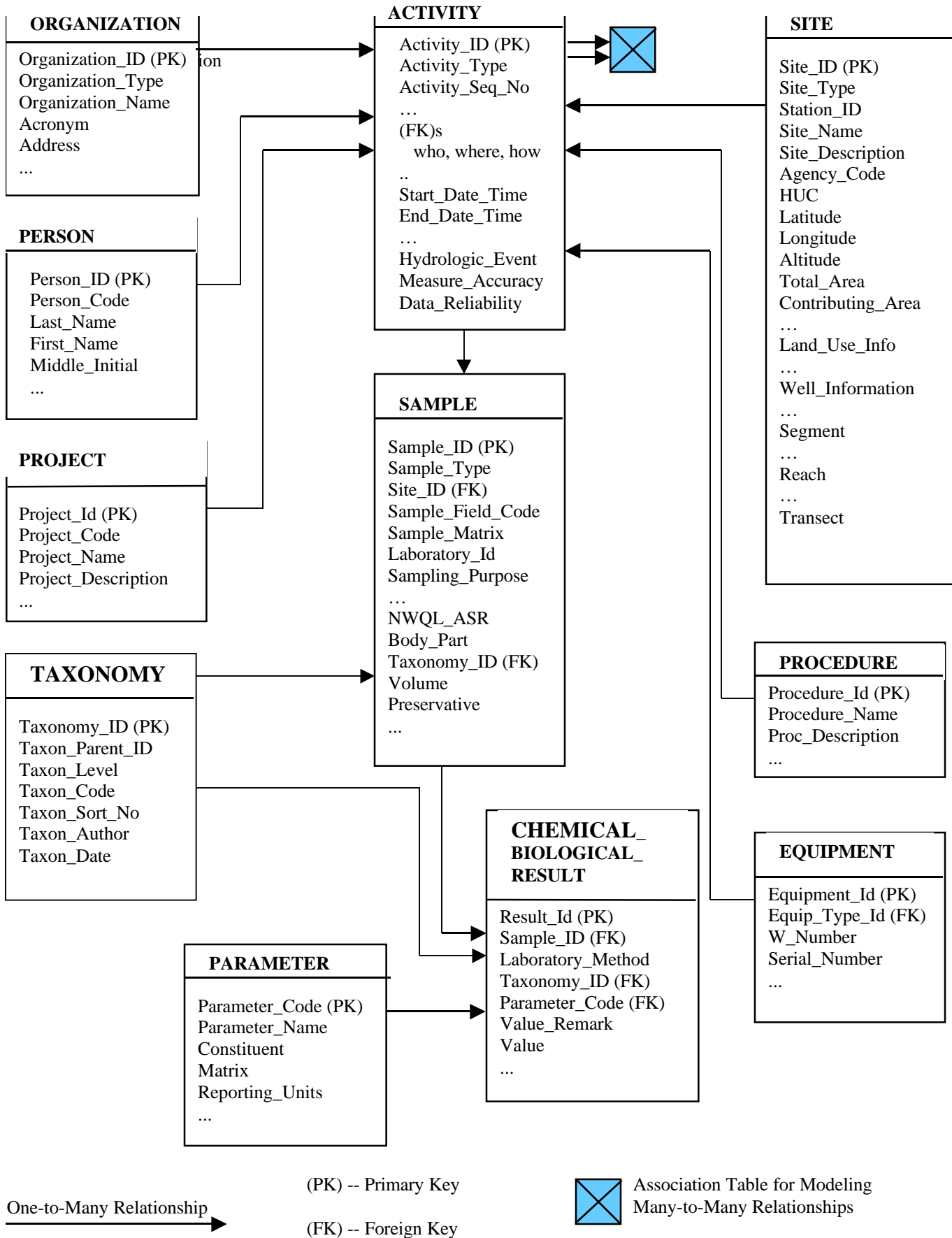
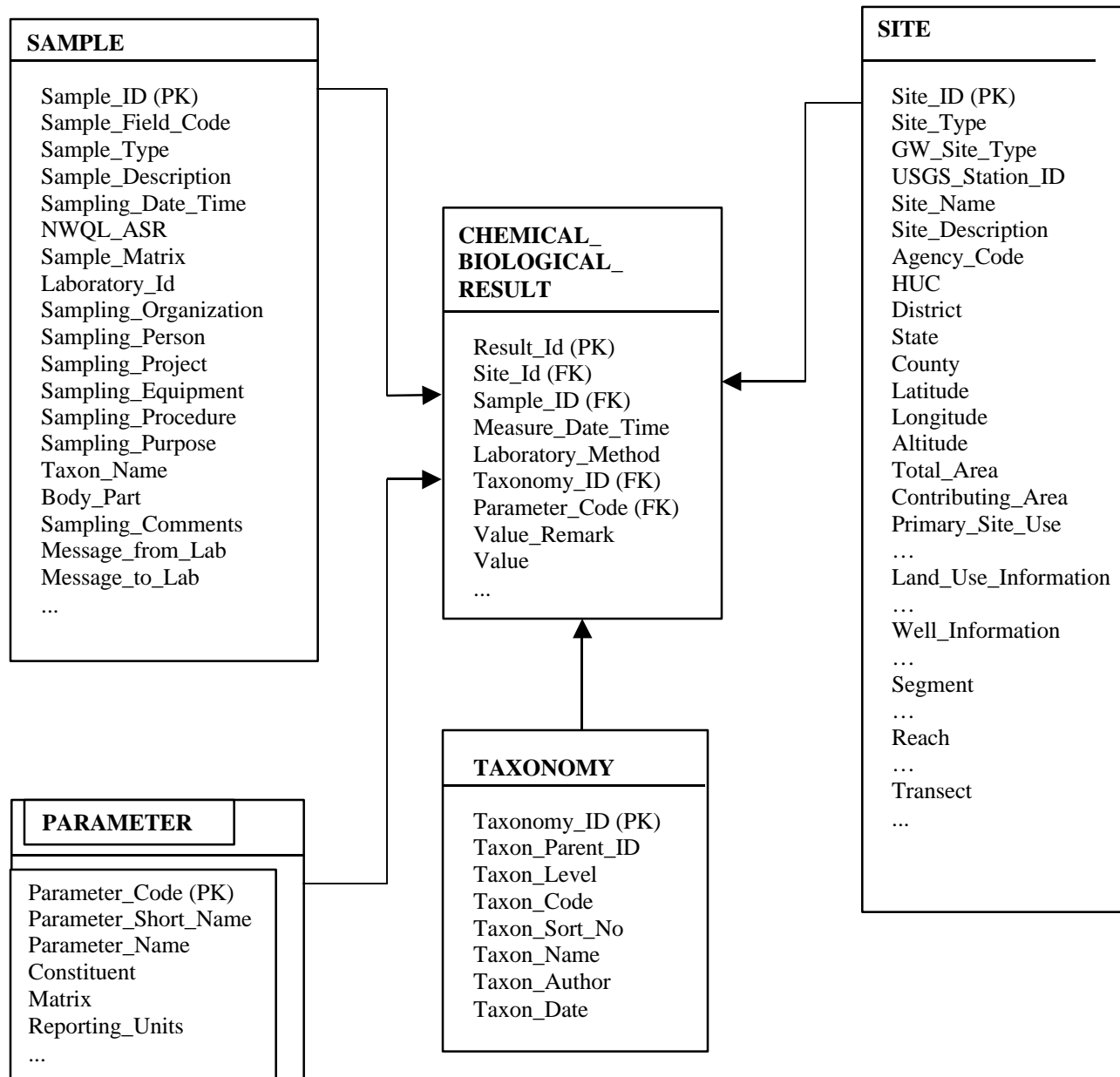


Fig 3 Schematic Design of the Ecological Synthesis Transactional Relational Database



One-to-Many Relationship →

(PK) -- Primary Key

(FK) -- Foreign Key

Fig. 4 Schematic “Star Schema” of Ecological Synthesis Data